基于二部图多权重投影的大数据推荐算法 *

高 薇[1]，何可期[2]

(1. 闽南理工学院，信息管理学院，福建 石狮 362700; 2. 中山大学，大数据与计算机学院，广州 510275)

摘 要：基于网络结构的推荐算法存在多样性不足的问题，为此提出了一种二部图多权重投影的大数据推荐算法。首先，提取出数据集的基础信息，将所有的项目—用户数据输入莱文斯坦距离程序，计算各个属性之间的相似性；然后，计算二部图网络中节点之间相同邻居的数量、节点之间的共同邻居度以及每个节点的度，计算二部图网络中每条边的三重权重；最后，采用增强的二部图投影技术提取二部图网络的潜在链接，实现基于相似性的链接预测。采用大数据集与小数据集分别完成了实验，结果显示该算法的准确率与覆盖率均优于其他几种类型的推荐算法，并且优于同类型的推荐算法。

关键词：推荐系统；大数据技术；二部图网络；链接预测；网络投影；单模网络

中图分类号：TP391 doi: 10.3969/j.issn.1001-3695.2018.07.0612

Recommendation system of big data based on multi-weight projection of bipartite network

Gao Wei[1], He Keqi[2]

(1. *School of Information Management*, *Minnan Institute of Technology*, *Shishi Fujian* 362700, *China*; 2. *Institute of Big Data & Computer*, *Sun Yat-sen University*, *Guangzhou* 510275, *China*)

Abstract: Most recommendation systems based on the network structure suffer from lack of diversity, so that a recommendation system of big data based on multi-weight projection of bipartite network is proposed. Firstly, the basic information of datasets is abstracted, items-users lists are applied as an input to Levenshtein Distance algorithm to compute similarity of each property; then, the number of common neighbors of the nodes in the bipartite network, the degree of common neighbors of the nodes in the bipartite network and degree of each node in bipartite network are all computed, triple weights of each side of the bipartite network are computed; lastly, the enhanced bipartite projection technique is adopted to abstract the potential links of the bipartite network to realize the link prediction based on similarity. The experiments based on both of big dataset and small dataset are realized, the results show that the proposed algorithm outperforms different kinds of recommendation systems in terms of accuracy and coverage of recommendation, at the same time, it outperforms the other recommendation system based on network structure.

Key words: recommendation system; big data technique; bipartite network; linkage prediction; network projection; one-mode network

0 引言

互联网的普及使得网络中的数据量急剧增长，许多大型网站日均访问量巨大，并且包含大量的项目信息，如京东商城[1]、淘宝网[2]、豆瓣网[3]、知网、网易云音乐等。大量的冗余信息极大地降低了用户的检索效率，不仅影响了用户的满意度，也为门户网站带来了巨大的负担。个性化推荐系统是解决上述问题的一个重要方案[4]，根据用户的购买记录、评论信息以及评分信息为用户推送合适的项目，降低用户的访问时间，模拟销售人员的推荐效果。

当前主流的推荐系统主要分为协同过滤推荐算法、基于内容的推荐算法、组合推荐算法、基于网络结构的推荐算法四类[5, 6]。协同过滤推荐算法易于实现，使用最为广泛[7]，基于内容的推荐算法根据用户过去的喜好推荐类似的项目[8]。这两种算法易于实现，也可以获得较高的推荐准确率，但是在多样性方面略有不足，并且受到冷启动问题的限制。组合推荐算法则一般结合两个互补的推荐算法，同时提高推荐系统的推荐准确率与推荐覆盖率两个重要指标，但组合推荐算法的计算效率较低，对于大数据的推送实时性不足[9]。基于网络结构的推荐算法是近期受到关注的一类方案，该方案不仅获得了较好的推荐准确率，而且也实现了较高的计算效率[10]。

许多现实问题可建模为一个网络结构，网络的节点表示问题的各个实体，网络的边表示实体之间的关系。二部图网络可描述复杂且规模庞大的问题[11]，包括社交网络、电子商务、生物信息领域等。目前二部图网络已在推荐系统问题上取得了一定的成效，主要通过二部图建模项目与用户群体，然后将项目与用户之间的关系建模为链接，通过已有的链接预测网络的潜在链接，这些潜在链接即为用户可能偏爱的项目。大多数基于二部图网络的推荐算法将流行的项目推送给用户，为用户推荐冷门项目的数量则明显不足，对推荐的多样性具有不利的影响[12]。引起推荐多样性不足的主要原因在于，在实际数据集建模网络的过程中，仅仅考虑了数据集内的强关系，忽略了弱关系和隐藏信息，虽然该机制提高了算

收稿日期：2018-07-23；**修回日期：**2018-10-26 **基金项目：**国家自然科学基金资助项目（61471161）；福建省教育厅 2015 年高等学校创新创业教育改革立项项目（闽教高〔2015〕41 号）

作者简介：高薇（1981-），女，吉林人，讲师，硕士，主要研究方向为计算机科学与技术、人工智能、大数据算法等(zhumeigg@126.com)；何可期（1988-），男，湖北武汉人，博士研究生，主要研究方向为高性能计算、算法研究等.

法的处理效率，但是牺牲了算法的多样性。

为了解决上述问题，通过投影将二部图网络转换为单模网络，采用增强的加权单模投影网络保留骨干网络，过滤原数据集的冗余信息，同时保留网络的强、弱关系信息。在过滤冗余信息的过程中，降低了投影网络的信息量，从而提高链接预测的计算效率。在投影网络的链接中考虑了邻居数量、共同邻居度以及节点的度三重关系，保留了原数据集的强关系与弱关系。

1　预测二部图网络的链接

一个二部图可表示一个网络，其节点分为两个不相交集合，设为 U 与 V，一个 U 节点与一个 V 节点之间存在一条链接。一个二部图网络定义为 $G=(U, V, E)$，U 与 V 是两个不相交节点集，E 是二部图网络 G 的边集合，$L(U)$是网络 G 中节点 U 的邻居节点集，$L(V)$ 是网络 G 中节点 V 的邻居节点集。在二部图网络中，同一个节点集内的节点没有链接。

定义 1　二部图网络。二部图网络表示为一个三元组形式 $G=(U,V,E)$，其中 U 与 V 是 G 的两个节点集，$E\subseteq|U|\times|V|$ 是 G 的边集合。相同节点集内的节点之间没有连接。

将二部图网络表示为$|U|\times|V|$的矩阵形式，U 中共有 n 个节点，V 中共有 m 个节点，网络 G 可表示为 $m\times n$ 维的邻接矩阵。二部图网络 G 的矩阵元素 A_{ij} 与对角矩阵 A 分别定义为

$$A_{ij}=\begin{cases}1, & if\,(u_i,v_j)\in E\\ 0, & \text{其他情况}\end{cases},\quad A=\begin{bmatrix}0_{m\times n} & A_{n\times m}\\ A^T_{m\times m} & 0_{n\times m}\end{bmatrix}\tag{1}$$

其中：$0_{n\times n}$ 与 $0_{m\times n}$ 分别为 $n\times n$ 与 $m\times n$ 的全零矩阵；$A_{n\times m}$ 为非零矩阵。所以邻接矩阵具有对称性，使用 $A_{n\times m}$ 矩阵表示二部图网络 G，U 集合的每行与每列表示 V 集合的一个节点。

二部图网络链接预测问题的目标是寻找网络当前不存在但未来会出现的链接。假设 $G=(U,V,E)$是时间 t 的二部图，链接预测任务是预测时间 t+1 二部图网络中的新链接。

将二部图转换为单模网络是分析二部图网络广泛使用的一个方案。投影技术是将二部图网络转换为单模网络的有效技术，投影后的网络是典型的单模网络结构。为了预测网络的潜在链接，首先将二部图网络转换为投影网络。投影网络如下定义：

定义 2　投影网络。$G=(U,V,E)$是一个二部图网络，并且$|U(G)|=m$，$|V(G)|=n$，$|E(G)|=m\times n$ 维。二部图网络两个节点集 U 与 V 分别转换为两个投影网络，获得 U-投影网络 $G_u=(U, E_u)$与 V-投影网络 $G_v=(U, E_v)$。

$$E_u=\{(u_i,u_j)|u_i,u_j\in U,|\exists v_i\in V,\ v_i\in\Gamma(u_i)\cap\Gamma(u_j)\},$$

$$E_v=\{(v_i,v_j)|v_i,v_j\in V,|\exists u_i\in U,\ u_i\in\Gamma(v_i)\cap\Gamma(v_j)\}\tag{2}$$

根据定义 2，$u_i v_e u_j$ 节点是二部图网络中 U 集合的元素。如果二部图网络中 V 节点有一个以上的邻居节点，那么 U-投影网络(G_u)中节点 u_i 与 u_j 之间存在一个链接。与之相似，G 的 V-投影网络定义为 $G_v=(U, E_v)$。

二部图网络转换为单模网络之后，原网络的拓扑结构信息可能丢失。为了解决该问题，设计了加权的二部图网络。加权二部图网络可以表示网络的拓扑属性，如项目—用户推荐关系、作者—文献关系、病人—病情关系等。加权的单模网络投影技术从一个二维网络获得加权的单模网络，其中边的权重表示节点的共同邻居数量。使用一个加权单模投影获得的网络称为加权投影网络，其数学模型定义为：

定义 3　加权投影网络。$G=(U,V,E)$是一个二部图网络，其空间维度为$|U(G)|=m$，$|U(G)|=n$，$|E(G)|=m\times n$。二部图网络的 U-投影网络 $G_u=(U, E_u)$，V-投影网络 $G_v=(U, E_v)$。为每个 E_u 与 E_v 边进行加权处理，加权函数 W 定义如下：

$$\begin{aligned}&E_u=\{(u_i,u_j)|u_i,u_j\in U,|\exists v_i\in V,\quad v_i\in\Gamma(u_i)\cap\Gamma(u_j)\},\\&W:(u_i,u_j)\rightarrow|\Gamma(u_i)\cap\Gamma(u_j)|\\&E_v=\{(v_i,v_j)|v_i,v_j\in V,|\exists u_i\in V,\ u_i\in\Gamma(v_i)\cap\Gamma(v_j)\},\\&W:(v_i,v_j)\rightarrow|\Gamma(v_i)\cap\Gamma(v_j)|\end{aligned}\tag{3}$$

其中：$W(u_i, u_j)$分别表示 i 节点与 j 节点的重要性。对于 U-投影网络，$W(u_i, u_j)$定义为 $\Gamma(i)\cap\Gamma(j)$相同邻居节点的数量。

因为从实际数据构建网络，网络中包含大量的冗余信息，这些冗余信息对潜在链接的预测具有不利的影响，所以首先需要过滤数据集的冗余链接，可有效地降低网络的复杂度，并且提高链接的预测质量。本文设计了骨干网提取算法将加权单模网络转换为增强的单模网络，增强的单模网络如下定义：

定义 4　增强的二部图投影。U、V 两个节点集和链接集 E 组成了二部图网络 $G=(U,V,E)$，从二部图网络 G 获得两个投影网络 $G_u=(U, E_u)$与 $G_v=(U, E_v)$。如果 A 与 B 边的权重 $W(A,B)>\alpha$，那么(A,B)边划分为增强网络 SBP^{α}；反之，(A,B)边属于冗余信息，过滤这些边。

2　本方案的主要内容

本文从大规模实际数据集构建了复杂的二部图网络，设计了二部图网络的链接预测方案，图 1 所示是本方案的流程框图,其中包括了基本信息提取方案，另一个关键工作是根据当前的二部图网络预测网络的潜在链接。项目信息、用户、用户评价等数据组成了二部图网络，本文的关键工作是从二部图网络提取基本信息。

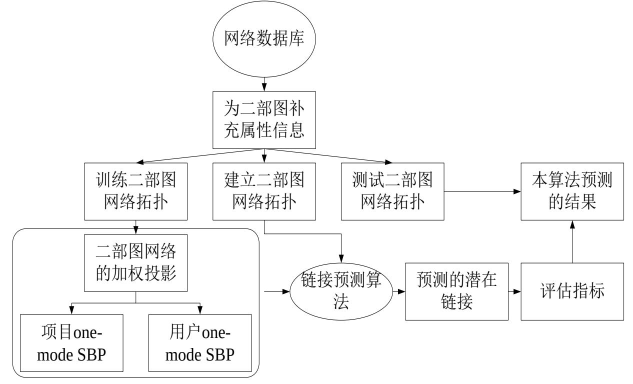


图 1　本方案的流程框图

Fig. 1　Block diagram of proposed schema

2.1　提取基础信息

图 2 所示是基础信息的提取流程框图，提取了项目的详细信息与用户信息。将所有的项目—用户数据输入莱文斯坦距离算法(levenshtein distance algorithm, LDA)，LDA 算法评估两个词汇之间的相似性。首先，分析原数据集，创建每个项目的名称列表，为选择同一个项目名称的用户建立关联性，并且保持用户的 ID 信息与日期信息；然后，合并数据集的项目标签，将所有的项目标签与用户分别建立新的列表，将这两个列表输入 LDA 算法重新计算；最后，将标签列表内的标签标准化处理，根据标准化的标签更新项目—用户链接信息。例如电影项目的名称为“卧虎藏龙”，标签为“动作片”，将网络中电影项目的标签统一标准化处理。

2.2　链接预测算法

采用增强的二部图投影技术提取二部图网络的潜在链接，实现基于相似性的链接预测。首先，使用挖掘技术从大

数据集中提取二部图网络(如定义 1)，采用 2.1 节的方案提取数据集的相关信息；然后，二部图网络转换为加权的单模网络(如定义 3)，将单模网络转换为投影网络的骨干网络 SBP^{α}(如定义 4)。

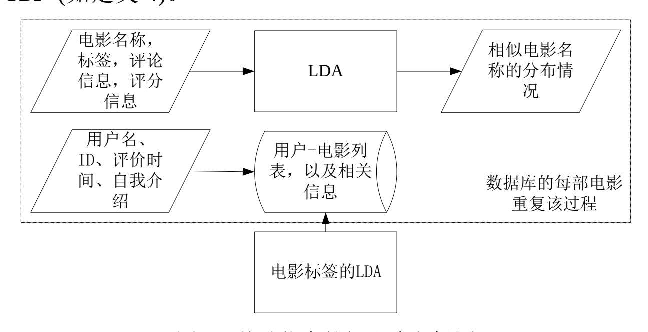


图 2　基础信息的提取流程框图

Fig. 2　Block diagram of basic information abstraction

在 SBP^{α} 单模网络中，检测具有内部节点对属性的节点对，预测网络的潜在链接。该方案降低了预测的链接数量，同时提高了预测的质量。如果二部图网络 G 中两个子节点之间存在交互，那么这两个子节点可能存在潜在链接；如果两个子节点没有共同的邻居节点，那么这两个子节点存在潜在链接的概率较低。

定义 5　潜在链接(potential links, PL)。假设 $G=(U,V,E)$ 是一个二部图网络，$G_u^{\alpha}=(U,E_u^{\alpha})$ 是 U-投影单模网络，$G_v^{\alpha}=(U,E_v^{\alpha})$ 是 V-投影单模网络，式中 α 表示单模网络。假设 $A\epsilon U$ 是 U 投影网络 $G_u^{\alpha}=(U,E_u^{\alpha})$ 的节点，那么节点 A 具有一个潜在链接的概率：

$PL_A=\{K\backslash(\Gamma_A)_{train}\}$，$K=\Gamma(k_1)\cup\Gamma(k_2),\ldots,\cup\Gamma(k_n)$表示 G_{train} 网络中一个节点 Γ_A^{α} 的邻居，节点 A 属于投影网络 G_u^{α}，$\Gamma(k_n)$ 表示节点 k_n 在二部图 G 中的邻居节点。如果满足式(1)，则 $PL_A=\{p|p\in K\wedge p\notin\Gamma_{A'}(A,p)\}$，$PL=\{PL_A\cup PL_B,\ldots,\cup PL_m\}$ 表示预测的潜在链接。

存在潜在链接的节点应当满足

$$\Gamma_U(A)\cap\Gamma(p_i)\neq\varnothing \text{ 并且 } p_i\notin\Gamma(A) \tag{4}$$

其中：Ø 表示空集；$A\in U$ 与 $p_i\in V$ 表示两个节点，并且$(A, p_i)\notin E$。

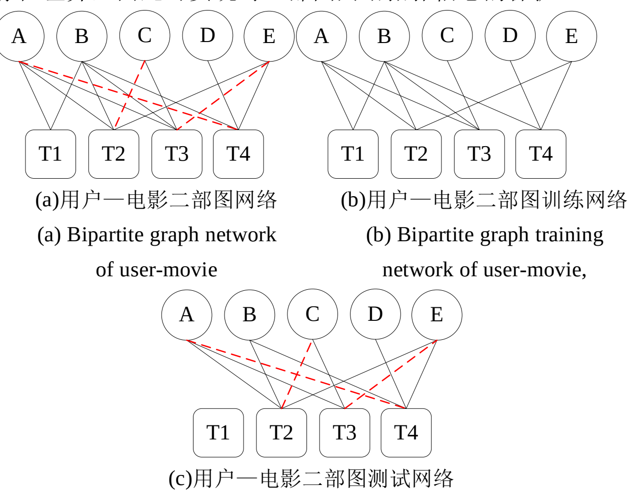
定义 6　潜在链接的覆盖模式。假设 G_{train} 为二部图训练网络，G_u^{α} 为 U-投影单模网络。每个 $C\epsilon(\Gamma_U(A))_{train}\cap\Gamma(p_i)$的潜在链接是 PL(潜在链接)覆盖的模式。PL(潜在链接)覆盖的模式数量越多，那么潜在链接变为真实链接的概率越高。因此，PL 覆盖的模式数量越多可用于评估潜在链接的概率。

例如图 3(a)所示的二部图网络中，圆形节点表示用户，方形节点表示用户观看的电影。T_2、T_4 是用户 B 与 E 共同观看的电影，(E,T_1)与(E,T_3)是满足定义 5 条件的潜在链接。这两个 PL 覆盖的模式数量不同，(E,T_1)链接覆盖的模式为$\{E,A\},\{E,B\}$；(E,T_3)链接覆盖的模式为$\{E,A\},\{E,B\},\{E,C\}$。(E,T_3)潜在链接的概率高于(E,T_1)潜在链接。

定义 7　改进的模式权重。PL 覆盖的模式数量越多，网络中 PL 覆盖的每个链接权重则越重要[13]。计算$\{A, B\}$的模式权重，需要考虑三个基本因素：

a)　二部图网络中 A、B 节点共同的邻居数量。

潜在链接覆盖的模式等于投影网络的边。二部图网络 G 中边 A 与 B 有相同的邻居节点，表示为投影网络 G_u 的(A,B)边。图 4(a)与(c)是两个不同的二部图网络，但它们的投影网络相同，如图 4(b)(d)所示。这种情况下，投影网络丢失了二部图网络的拓扑信息，通过采用加权的二部图网络可维护网络的拓扑信息，图 4(a)的邻居数量小于(b)，所以(A, B)边的值存在差异，由此可实现对二部图网络拓扑信息的保护。



(a)用户—电影二部图网络　(b)用户—电影二部图训练网络

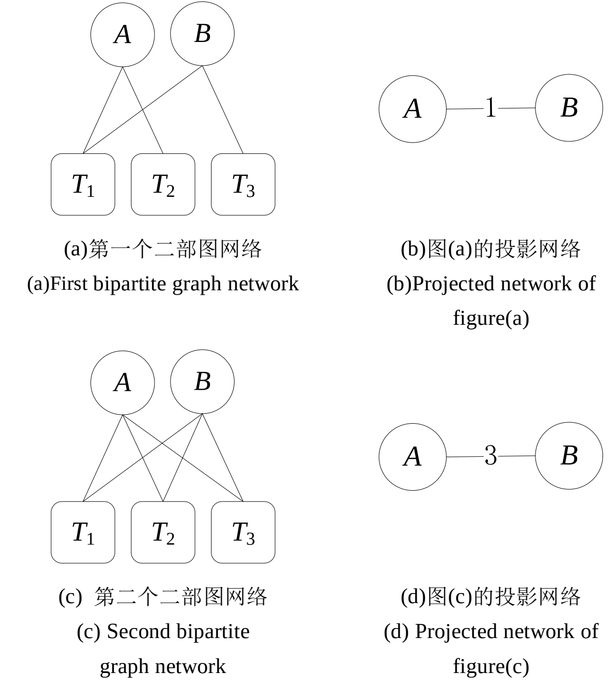
(a) Bipartite graph network of user-movie　(b) Bipartite graph training network of user-movie,

(c)用户—电影二部图测试网络

(c)Bipartite graph testing network of user-movie

图 3 潜在链接覆盖模式的实例

Fig. 3　Case of potential linkage coverage model



(a)第一个二部图网络　(b)图(a)的投影网络

(a)First bipartite graph network　(b)Projected network of figure(a)

(c) 第二个二部图网络　(d)图(c)的投影网络

(c) Second bipartite graph network　(d) Projected network of figure(c)

图 4 两个不同的二部图网络以及加权投影网络

Fig. 4　Two different bipartite graph networks and weighted projected network

b)　二部图网络中节点 A 与 B 的共同邻居度。

二部图网络中共同邻居的度也是重要的隐藏信息。如果二部图网络中两个节点共同邻居的度增加，那么两个节点的相似性也提高。例如图 5(a)中 T_2 的度是 A、B 用户的共同观看的电影，(b)中 A、B、C、D 用户观看了电影 T_2，(a)和(b)的共同邻居度分别为 2 与 4。两个节点的共同邻居度越小，那么两者的相似性越高，模式的权重也越高。

c)　二部图网络中节点 A 与 B 的度。

A 与 B 节点的度对于$\{A,B\}$边的模式权重具有一定的影响。例如：图 5(a)中节点 A 与 B 的度分别为 2 与 1，(b)中节点 A 与 B 的度分别为 2 与 3。在图 5(a)中，当用户 A 观看电影 T_1 与 T_2；用户 B 仅观看电影 T_2。而在图 5(b)中 A、B 观看了不同的电影，他们观看的共同电影是 T_2，图 5(b)中两者的相似性低于图 5(a)，A 与 B 权重也应当高于图 5(a)。

假设 $G_u=(U, E_u)$是从二部图网络 $G=(U, V, E)$获得的投影

单模加权网络，$(A, B)\epsilon E_u$ 是网络 G_u 内的一条边。$\{A,B\}$模式的权重 $w(A, B)$计算为

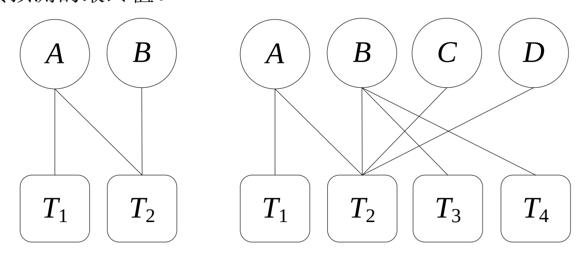
$$w(A,B)=\frac{2}{k_A+k_B}\sum_{z\in\Gamma(A)\cap\Gamma(B)}\frac{1}{k_z} \tag{5}$$

其中：k_a、k_b、k_z 分别为节点 A、B、C 在二部图网络 G 中的度；$T(A)$、$T(B)$分别为二部图网络中 A 与 B 节点的邻居集合。

从式(5)可看出，A、B 节点的共同邻居度越小，模式的权重越大。PL 覆盖的每个模式元素概率等价于 PL 覆盖的模式总权重，(A, p_i)潜在链接的最终总评分计算为

$$S(A,p_i)=\sum_{A,B\in\Gamma(A,p_i)} w(A,B) \tag{6}$$

从式(6)可看出，高权重模式的潜在链接越多，则链接预测的概率越高。因此，潜在链接覆盖的模式权重之和即为潜在链接预测的最终值。



(a) 4 个节点的二部图网络　(b) 8 个节点的二部图网络

(a)Bipartite graph network of 4 nodes　(b) Bipartite graph network of 8 nodes

图 5　两个不同的二部图网络

Fig. 5　Two different bipartite graph networks

2.3　算法的具体实现

采用文献[14]的二部图网络的内部链接定义，基于该定义设计了本文的链接预测方案，算法 1 所示是基于二部图网络链接预测的推荐算法伪代码。文献[13]基于内部链接定义了候选节点对，基于此概念提出一个链接预测方案，该算法对大规模网络的效率较低。本文则提出了适合大规模二部图网络的链接预测方案。

本文从原数据集获得了包含弱关系的投影单模网络，该网络过滤了冗余的信息，提取了骨干网络。根据预定义的阈值创建一个增强的骨干网络，在初始化的单模投影网络中，根据高频率边权重决定该阈值。该方案有效地缩小了潜在链接的集合，降低了算法的总体计算时间，同时也维护了网络中强关系的节点。

算法 1　基于二部图网络链接预测的推荐算法

输入：二部图网络 $G(U, V, E)$，训练二部图网络 $G_{train}(U, V, E_{train})$，阈值 α。

输出：潜在链接集合 PL，PL 中各元素的最终评分。

/*模块 1：根据定义 3 构建加权的投影网络 G_u*/

for each node A in U

　for each x in $\Gamma(A)$　/*$\Gamma(A)$表示 A 的邻居节点*/

　　for each B in $\Gamma(x)$　/*$\Gamma(x)$表示 x 的邻居节点*/

　　　W：$(A, B)\rightarrow|\Gamma(A)\cap\Gamma(B)|$/*计算 A 与 B 的共同邻居度*/

　　　$E_u=E_u\cup\{(A, B)\}$

　　end for

　end for

end for

/*模块 2：根据定义 4 构建强化的投影网络 G_u^α */

for each A in G_u

　for each B in $\Gamma_u(A)$　/*式中 B 为 A 的邻居节点*/

　　if $W(A, B) > \alpha$ then

　　　$E_u^\alpha = E_u^\alpha \cup\{(A,B)\}$；　/*根据定义 4 建立投影*/

　　end if

　end for

end for

/*模块 3：根据定义 5 构建 PL 集合*/

for each 节点 A in G_u^α　/* G_u^α 是强化的投影网络*/

　for each A 的邻居节点 K　/*遍历 A 的每个邻居节点*/

　　for each 与 K 连接的节点 p　/*遍历 K 直接连接的节点*/

　　　if $(A,p)\notin E_{train}$ then

　　　　$PL_A=PL_A\cup(A, p)$；/*将 PL_A 边的权重累积起来*/

　　　end if

　　end for

　end for

　$PL=PL\cup PL_A$

end for

/*模块 4：计算 PL 集合每个元素的最终评分*/

for each 节点对(A, p) in PL

　for each 节点 $C(C\in(\Gamma u(A))_{train}\cap\Gamma(p))$　/*遍历每个节点*/

　　根据式(2)计算节点对(A, p)的权重；

　　$S(A, p)= S(A, p)+w(A, C)$；　/*即式(3)*/

　end for

end for

3　实验与结果分析

为了评估基于二部图网络链接预测的推荐算法性能，完成了多组实验，通过推荐准确率与推荐覆盖率两个指标评估推荐算法的性能。实验环境为 PC 机：Intel Core i7 处理器，Windows 10 操作系统，12 GB 内存。

3.1　实验数据集

实验采用 FilmTrust 与 Epinions 两个数据集，表 1 所示是两个数据集的基本介绍。FilmTrust 是从 FilmTrust 网站采集的小数据集[15]，该数据集共包含 35 497 个评分，1 642 个用户，2 071 部电影。评分范围为 0.5,1.0,1.5,…,4.0，同时包含了 1 853 条用户评论。第二个数据集是 Epinions 大数据集，该数据集由 Paolo Massa 从 Epinions.com 网站收集，该数据集的评分范围为 1,2,3,4,5，该数据集包含 49 290 个用户，139 728 个项目，实验从 Epinions 数据集中随机选择了 5 000 个用户与 10 000 个项目作为实验数据集。

表 1　FilmTrust 与 Epinions 两个数据集的基本信息

Table 1　Basic information of FilmTrust and Epinions datasets

数据集	FilmTrust	Epinions
用户数量	1 508	5 000
项目数量	2 071	10 000
评分数量	35 500	77 100
评分粒度	0.5	1
评分范围	[0.5, 4]	[1, 5]
评论数量	1 853	20 500

3.2　性能评价指标

采用平均绝对误差(mean absolute deviation, MAE)指标评估本方案的推荐准确率，对于一个包含 N 个评分的数据集，MAE 的计算方案为

$$MAE=\frac{\sum_{i=1}^{N}|r_i-r_p|}{N} \tag{7}$$

其中：r_p 为目标项目 i 的预测评分；r_i 是 i 的实际评分。MAE 值越低，推荐准确率越高。覆盖率是评估推荐系统性能的另

一个重要指标，计算式为

$$RC=M/|\Omega| \tag{8}$$

其中：M 表示预测评分的数量；$|\Omega|$表示数据集内的评分总数量。RC 值越高表示推荐系统的覆盖性能越好。F_1 指标评估推荐系统的总体性能为

$$F1=\frac{2\times precision\times RC}{precision+RC} \tag{9}$$

其中：精度 $precision$ 定义为

$$\text{precision}=1-\text{MAE}/(r_{\max}-r_{\min}) \tag{10}$$

其中：r_{max} 与 r_{min} 分别表示推荐系统内最高评分值与最低评分值，分别为 1 与 0。

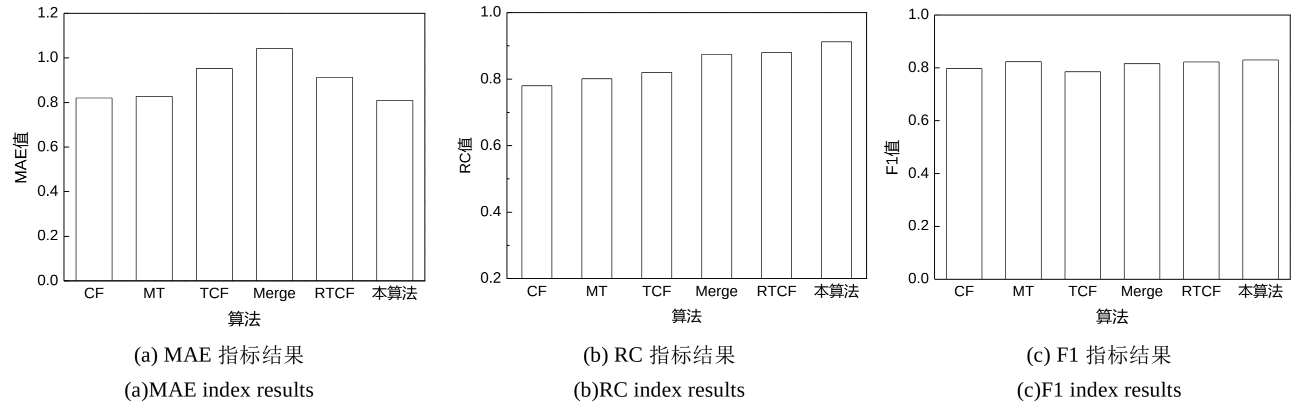
3.3 实验结果与分析

本算法是一种基于二部图网络的推荐算法，选择其他不同类型的推荐算法与本算法比较，包括协同过滤推荐算法(collaborative filtering recommender system, CF)[16]、基于信任的可靠推荐算法(more trust-aware recommender system, MT)[17]、基于内容的推荐算法(term weights for content-based filtering recommender, TCF)[18]、协同过滤与社会关系的组合推荐算法(merging collaborative filtering and social relationships, Merge)[19]；此外还选择了两个近期的基于二部图推荐算法与本算法比较，分别为加权二部图推荐算法(real time construction full weighted bipartite, RTCF)[20]、二部图网络多视角分区的推荐算法(multiview bipartite network, MV)[21]。

3.3.1 正常状态的推荐实验

图 6 所示是 Epinions 数据集的推荐性能结果。图 6(a)~(c)分别是六种推荐算法的 MAE、RC 与 F1 指标结果。从结果可看出，本算法对于大规模数据集的推荐准确率与覆盖率均优于其他类型的推荐算法，并且也优于另一种基于二部图的推荐算法 RTCF。基于二部图的推荐算法在覆盖率上表现出一定的优势，主要在于 RTCF 与本算法均建立了较为全面的二部图，保留了数据集的诸多信息，而 CF、MT、TCF、Merge 四种算法在预处理的阶段，为了提高计算效率，过滤了数据集较多的信息。

图 7 所示是 FilmTrust 数据集的推荐性能结果。图 7(a)~(c)分别是五种推荐算法的 MAE、RC 与 F1 指标结果。从结果可看出，本算法对于小规模数据集的推荐准确率与覆盖率均优于其他类型的推荐算法，并且也优于另一个基于二部图的推荐算法 MV。综合两组实验的结果，本算法对小数据集的推荐准确率优于大数据集，原因在于本算法在提取二部图投影骨干网络的过程中，增加了冗余信息过滤的处理，该处理对大数据集删除的细节信息较多，影响了后期连接预测的准确率，导致对大数据集的推荐准确率降低，但是依然高于其他的推荐算法。

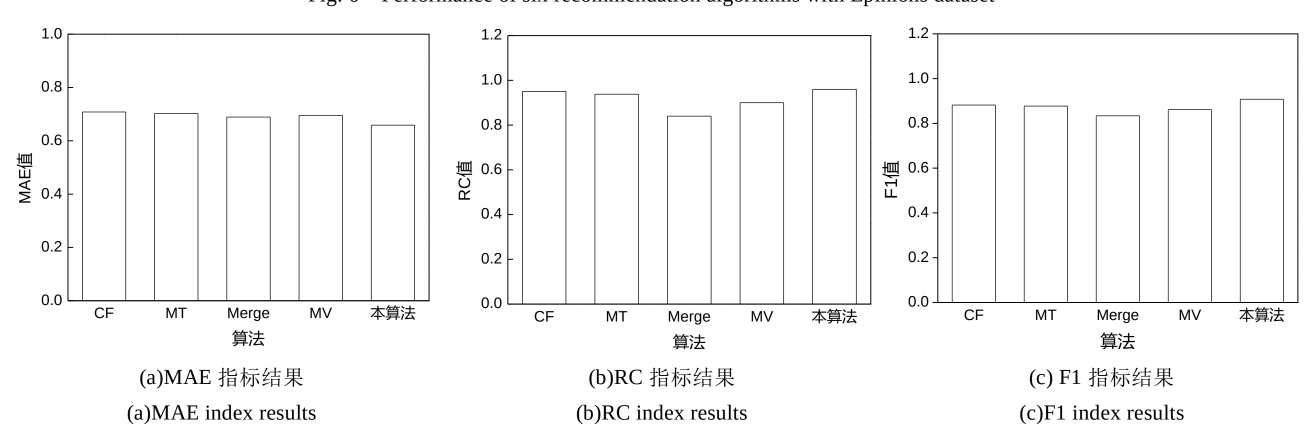


(a) MAE 指标结果 (b) RC 指标结果 (c) F1 指标结果

(a)MAE index results (b)RC index results (c)F1 index results

图 6 六种推荐算法对 Epinions 数据集的推荐性能

Fig. 6 Performance of six recommendation algorithms with Epinions dataset



(a)MAE 指标结果 (b)RC 指标结果 (c) F1 指标结果

(a)MAE index results (b)RC index results (c)F1 index results

图 7 五种推荐算法对 FilmTrust 数据集的推荐性能

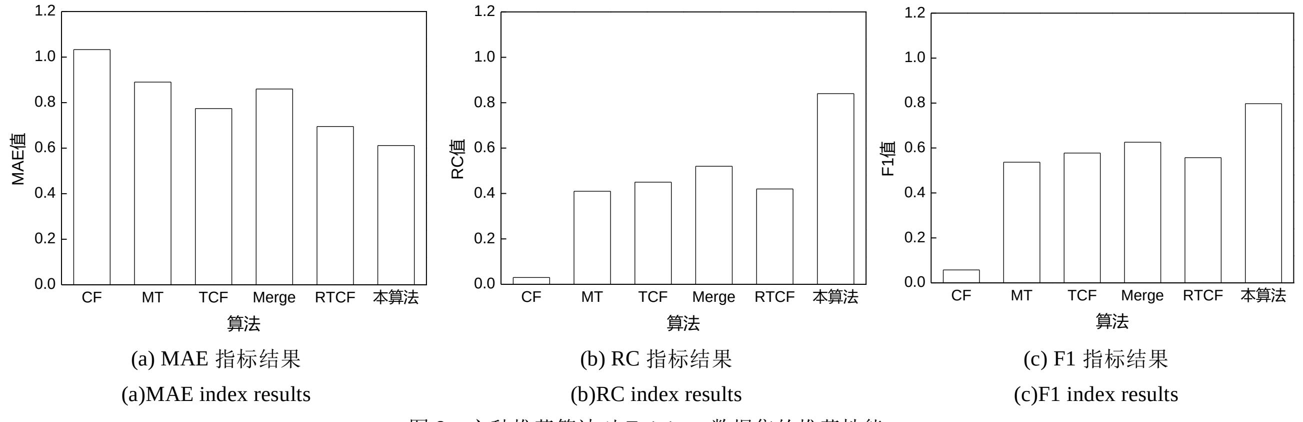
Fig. 7 Performance of five recommendation algorithms with filmtrust dataset

3.3.2 冷启动状态的推荐实验

冷启动是推荐系统的一个关键问题，是评价推荐系统的一个重要指标。因此对冷启动状态下的推荐系统也进行了实验，分析冷启动状态推荐系统的性能。图 8 所示是 Epinions 数据集的推荐性能结果。图 8(a)~(c)分别是六种推荐算法的 MAE、RC 与 F1 指标结果。从结果可看出，本算法对于大规模数据集的推荐准确率与覆盖率均优于其他类型的推荐算法，并且也优于另一种基于二部图的推荐算法 RTCF。协同过滤推荐系统受冷启动问题的影响较大，其推荐准确率、覆盖率均较差。MT 算法需要计算各个关系之间的信任值，受冷启动的影响较大，因此性能略差；TCF、Merge、RTCF 则表现出较为接近的性能，基于二部图的推荐算法受到冷启动

的影响较小，并且本算法通过三重权重建立投影网络的链接，能够有效地降低冷启动问题的影响。

图 9 所示是 FilmTrust 数据集的推荐性能结果。图 9(a)~(c)分别是五种推荐算法的 MAE、RC 与 F1 指标结果.从结果可看出，本算法对于小规模数据集的推荐准确率与覆盖率均优于其他类型的推荐算法，并且也优于另一种基于二部图的推荐算法 MV。协同过滤推荐系统受冷启动问题的影响较大，其推荐准确率、覆盖率均较差。MT 算法需要计算各个关系之间的信任值，受冷启动的影响较大，因此性能略差；Merge、MV 则表现出较为接近的性能。本算法的推荐准确率指标 MAE 仅为 0.26，获得了极高的准确率。本算法采用增强的加权单模投影网络保留骨干网络，过滤原数据集的冗余信息，同时保留网络的强、弱关系信息，该处理对于大数据集删除了较多的隐藏关系，但是对于小数据集则保留了丰富的强、弱关系信息，因此对小数据的推荐效果优于大数据集。

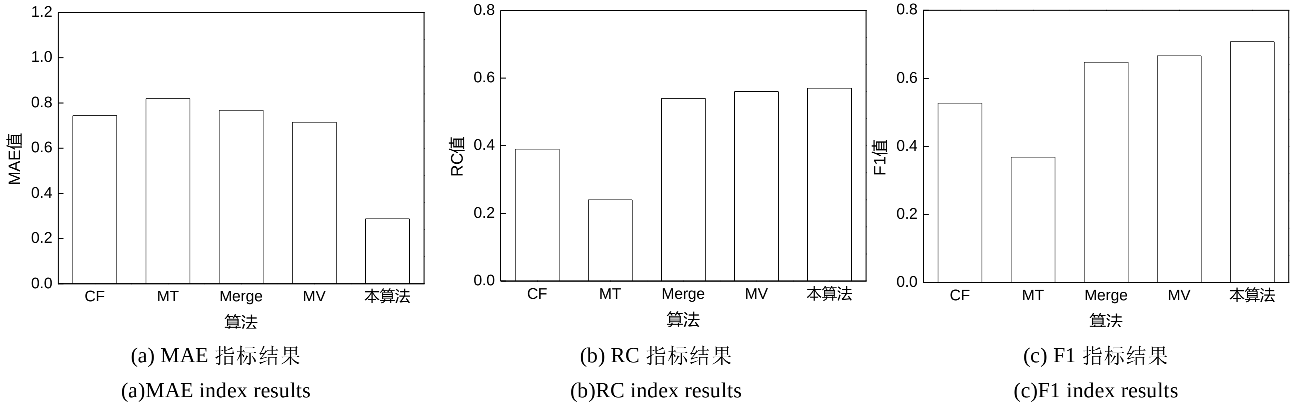


(a) MAE 指标结果　(b) RC 指标结果　(c) F1 指标结果

(a)MAE index results　(b)RC index results　(c)F1 index results

图 8　六种推荐算法对 Epinions 数据集的推荐性能

Fig. 8　Performance of six recommendation algorithms with Epinions dataset



(a) MAE 指标结果　(b) RC 指标结果　(c) F1 指标结果

(a)MAE index results　(b)RC index results　(c)F1 index results

图 9　五种推荐算法对 FilmTrust 数据集的推荐性能

Fig. 9　Performance of five recommendation algorithms with FilmTrust dataset

4　结束语

推荐系统在处理大数据时，在实际数据集建模网络的过程中，仅仅考虑了数据集内的强关系，忽略了弱关系和隐藏信息，虽然该机制提高了算法的处理效率，但是牺牲了算法的多样性。本文采用增强的加权单模投影网络保留骨干网络，过滤原数据集的冗余信息，同时保留网络的强、弱关系信息。本文对于大数据集与小数据集均进行了实验，并且考虑了冷启动的情况，结果显示本算法对于大小数据集均表现出一定的性能优势。

参考文献：

[1] 卿勇，刘梦娟，银盈，等. SMART: 一种面向电商平台快速消费品的图推荐算法 [J]. 计算机科学, 2017, 44 (b11): 464-469. (Qing Yong, Liu Mengjuan, Yin Ying, *et al.* SMART: a graph-based recommendation algorithm for fast moving consumer goods in E-commerce platform [J]. Computer Science, 2017, 44 (b11): 464-469.)

[2] 姚静天，王永利，侍秋艳. 基于联合物品搭配度的推荐算法框架 [J]. 上海理工大学学报, 2017, 39 (1): 42-50. (Yao Jingtian, Wang Yongli, Shi Qiuyan. Joint match degree of items for recommendation systems [J]. Journal of University of Shanghai for Science and Technology, 2017, 39 (1): 42-50.)

[3] 郭均鹏，赵梦楠. 面向在线社区用户的群体推荐算法研究 [J]. 计算机应用研究, 2014, 31 (3): 696-699. (Guo Junpeng, Zhao Mengnan. Group recommendation algorithm for online community users [J]. Application Research of Computers, 2014, 31 (3): 696-699.)

[4] 张时俊，王永恒. 基于矩阵分解的个性化推荐系统研究 [J]. 中文信息学报, 2017, 31 (3): 134-139. (Zhang Shijun, Wang Yongheng. personalized recommender system based on matrix factorization [J]. Journal of Chinese Information Processing, 2017, 31 (3): 134-139.)

[5] 刘华锋，景丽萍，于剑. 融合社交信息的矩阵分解推荐方法研究综述 [J]. 软件学报, 2018, 2 (2): 340-362. (Liu Huafeng, Jing Liping, Yu Jian. Survey of matrix factorization based recommendation methods by integrating social information [J]. Journal of Software, 2018, 2 (2): 340-362.)

[6] 张玉洁，杜雨露，孟祥武. 组推荐系统及其应用研究 [J]. 计算机学报, 2016, 39 (4): 745-764. (Zhang Yujie, Du Yulu, Meng Xiangwu. Research on group recommender systems and their applications [J]. Chinese Journal of Computers, 2016, 39 (4): 745-764.)

[7] Kumar V, Pujari A K, Sahu S K, *et al.* Collaborative filtering using

multiple binary maximum margin matrix factorizations [J]. Information Sciences, 2017, 380 (C): 1-11.

[8] 王光，张杰民，董帅含，等. 基于内容的加权粒度序列推荐算法 [J]. 计算机工程与科学, 2018, 1 (3): 564-570. (Wang Guang, Zhang Jiemin, Dong Shuaihan, *et al*. A content-based weighted granularity sequence recommendation algorithm [J]. Computer Engineering and Science, 2018, 1 (3): 564-570.)

[9] 伊华伟，张付志. 融合 k-距离和项目类别信息的鲁棒推荐算法 [J]. 小型微型计算机系统, 2017, 38 (11): 2476-2481. (Yi Huawei, Zhang Fuzhi. Robust Recommendation Method Fusing k-distance and Item Category Information [J]. Journal of Chinese Computer Systems, 2017, 38 (11): 2476-2481.)

[10] 李镇东，罗琦，施力力. 基于增加相似度系数的加权二部图推荐算法 [J]. 计算机科学, 2016, 43 (7): 259-264. (Li Zhendong, Luo Qi, Shi Lili. Weighted bipartite network recommendation algorithm based on increasing similarity coefficient [J]. Computer Science, 2016, 43 (7): 259-264.)

[11] Ge Mengqu, Li Ao, Wang Minghui. A biartite network-based method for prediction of lncRNA-protein interactions. [J]. Genomics Proteomics & Bioinformatics, 2016, 14 (1): 62-71.

[12] Daminelli S, Thomas J M, Durán C, *et al.* Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks [J]. New Journal of Physics, 2015, 17 (11): 113037.

[13] Cela K L, Sicilia M Á, Sánchez S. Social network analysis in E-learning environments: a preliminary systematic review [J]. Educational Psychology Review, 2015, 27 (1): 219-246.

[14] Wu Tsunghan, Yu Sheauharn, Liao Wanjiun, *et al.* Temporal bipartite projection and link prediction for online social networks [C]//Proc of IEEE International Conference on Big Data. 2015: 52-59.

[15] Guo Guibing, Zhang Jie, Thalmann D. Merging trust in collaborative filtering to alleviate data sparsity and cold start [J]. Knowledge-Based Systems, 2014, 57 (2): 57-68.

[16] Palomares I, Browne F, Wang Hui, *et al*. A Collaborative filtering recommender system model using owa and uninorm aggregation operators [C]// Proc of International Conference on Intelligent Systems and Knowledge Engineering. Piscataway, NJ: IEEE Press, 2016: 382-388.

[17] Abderrahim N, Benslimane S M. Towards improving recommender system: a social trust-aware approach [J]. International Journal of Modern Education & Computer Science, 2015, 7 (2): 8-15.

[18] Hardtke D, Hardtke D, Hardtke D, *et al.* Learning global term weights for content-based recommender systems [C]// Proc of International Conference on World Wide Web. Switzerland:International World Wide Web Conferences Steering Committee. 2016: 391-400.

[19] Meo P D, Ferrara E, Fiumara G, *et al.* Improving recommendation quality by merging collaborative filtering and social relationships [C]// Proc of International Conference on Intelligent Systems Design and Applications. Piscataway, NJ: IEEE Press, 2012: 587-592.

[20] Haihong E, Wang Jianfeng, Song Meina, *et al.* Incremental weighted bipartite algorithm for large-scale recommendation systems [J]. Turkish Journal of Electrical Engineering & Computer Sciences, 2016, 24 (2): 1-16.

[21] Tong Linqiao. Personal recommendation based on community partition of bipartite network [C]// Proc of International Conference on Cloud Computing and Big Data. Piscataway, NJ: IEEE Press, 2015: 336-341.